Youtube Analyzer: Scraping Youtube data from channels. Analysis of popular channels, type of content posted, frequency of posting, user engagement, etc.

ИЗРАБОТЕНО ОД:

РИСТОВА МАРИЈА

ИНДЕКС:

211126

Линк до google collab:

https://colab.research.google.com/drive/1kq_O7BHka_IoFxPacTgCtMpEYsmt1eL_?usp=sharing

Линк до видеото на youtube:

https://youtu.be/iBsAgbJYFW4

СОДРЖИНА

1. ВОВЕД

1.1 Цел на проектот

Овој проект е фокусиран на анализа на типовите на содржини на YouTube според насловите на видеата и бројот на лајкови за да се утврди кој тип на содржина е најпопуларен во конкретното видео и конкретниот канал. Во овој проект користејќи го YouTube Data API, собрани се податоци за видеа од популарниот канал “The Slow Mo Guys”, вклучувајќи ги насловите, датумите на објавување, опис, тагови, број на прегледи и бројот на лајкови. Со помош на машински модел за учење, насловите на видеата се класифицирани во различни категории а потоа резултатите се анализирани и визуелизирани за да се покаже кој тип на содржина има најголем број на лајкови за тој канал.

2. ПРИСТАП ДО ПРОБЛЕМОТ

2.1 Сеопфатен преглед

Проектот е поделен на неколку фази односно собирање на податоци, анализа на податоците и визуелизација на истите.

2.2 Фази на имплементација

1. Добивање на ID-а на видеата од YouTube каналот.
2. Собирање на податоци за видеата.
3. Класификација на видеата според насловот за кој тип на содржини се објавени.
4. Анализа на податоците за да се утврди кој тип на содржина има најмногу лајкови.

3. АЛАТКИ И БИБЛИОТЕКИ КОРИСТЕНИ

YouTube Data API – се користи за собирање на податоци за видеата, така што генериран е API клуч преку кој се собираат податоци од даден канал во овој случај насловите, датумите на објавување, бројот на прегледи, лајкови и коментари. Ова API овозможува да се добијат точни и ажурирани информации директно од YouTube платформата.

Pandas - се користи за манипулација и анализа на податоците. Pandas овозможува лесно ракување со табеларни податоци преку DataFrame објекти, кои се корисни за филтрирање, подредување и трансформирање на податоците. Со помош на Pandas, податоците се лесно подготвени за понатамошна анализа.

Transformers - модел од библиотеката Transformers кој е користен за класификација на насловите на видеата во овој случај во пет различни категории.

Matplotlib и Seaborn – се користени за визуелизација на резултатите. Со нивна помош, резултатите од анализата се прикажани на начин кој е лесно разбирлив и визуелно прикажан, што е од голема помош при презентирање на податоците.

4. ОБЈАСНУВАЊЕ НА КОДОТ

4.1 Добивање на видео ID-а

get_all_video_ids(channel_id) - Оваа функција се користи за добивање на сите видеа со нивните ID-а од YouTube каналот. Се креира празна листа video_ids за складирање на ID-а на видеата. Имаме променлива next_page_token која се користи за навигација низ страните. Во while циклисот има метод youtube.search.list() што прави барање до Youtube Data API. Параметрите part, channelId, type, maxResults и pageToken се зададени за да се специфицира што точно бараме и колку резултати да вратиме.

4.2 Собирање на податоци за видеа

get_video_info(video_ids) - се користи за добивање информации за даден сет на видеа. Се иницијализира нова празна листа која ги складира информациите за видеата. За секој item во response['items'] се извлекуваат информации и се додаваат во листата video_info, тие се претставени како:

- videoId: ID на видеото.
- title: наслов на видеото.
- description: опис на видеото.
- publishedAt: датум на објавување.
- tags: таговите на видеото.
- viewCount: број на прегледи.
- likeCount: број на лајкови.
- dislikeCount: број на дислајкови.
- commentCount: број на коментари.

4.3 Класификација на наслови на видеа

From transformers import pipeline делот од кодот се користи за да се вметне pipeline објектот од библиотеката Transformers, кој овозможува лесно користење на претходно тренирани модели. Се креира pipeline класификација користејќи го моделот BART-large-MNLI кој е веќе трениран на други податоци. Во овој случај зададов 4 лабели "Fashion", "Education", "Music" и "Gaming" и “Science”, по кои ги класифицирав видеата.

Пример. {'sequence': 'WHAT HAPPENS WHEN YOU DRINK 10 COFFEES IN ONE DAY',

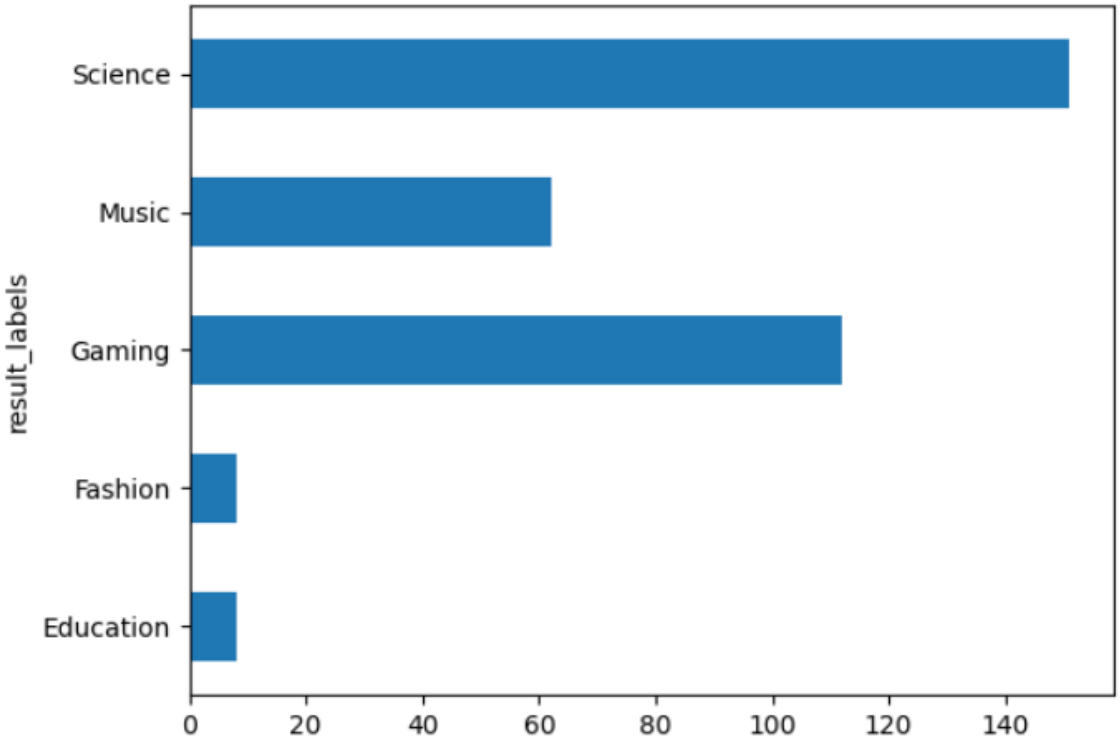
'labels': ['Science', 'Gaming', 'Education', 'Fashion', 'Music'],
'scores': [0.49128419160842896,
0.0007143329130485654,
0.0005000789533369243,
0.0004670500638894737,
0.00015751142927911133]},

Во scores се прикажани веројатностите за секоја категорија, таму каде што има најголема веројатност значи дека видеото спаѓа во таа категорија. Во овој случај

видеото припаѓа на категоријата Science, како што може да видиме од самиот наслов. Со многу помали веројатности припаѓа и во другите категории па затоа е квалификуван во делот Science.

5. РЕЗУЛТАТИ И ЗАКЛУЧОК

5.1 Резултати



Слика 1.

На сликата е прикажан столбест дијаграм кој ги визуелизира бројките за различните категории на видеа. Резултатите покажуваат дека категоријата „Science“ има најголем број, што укажува на висока популарност и интеракција на публиката со оваа категорија.

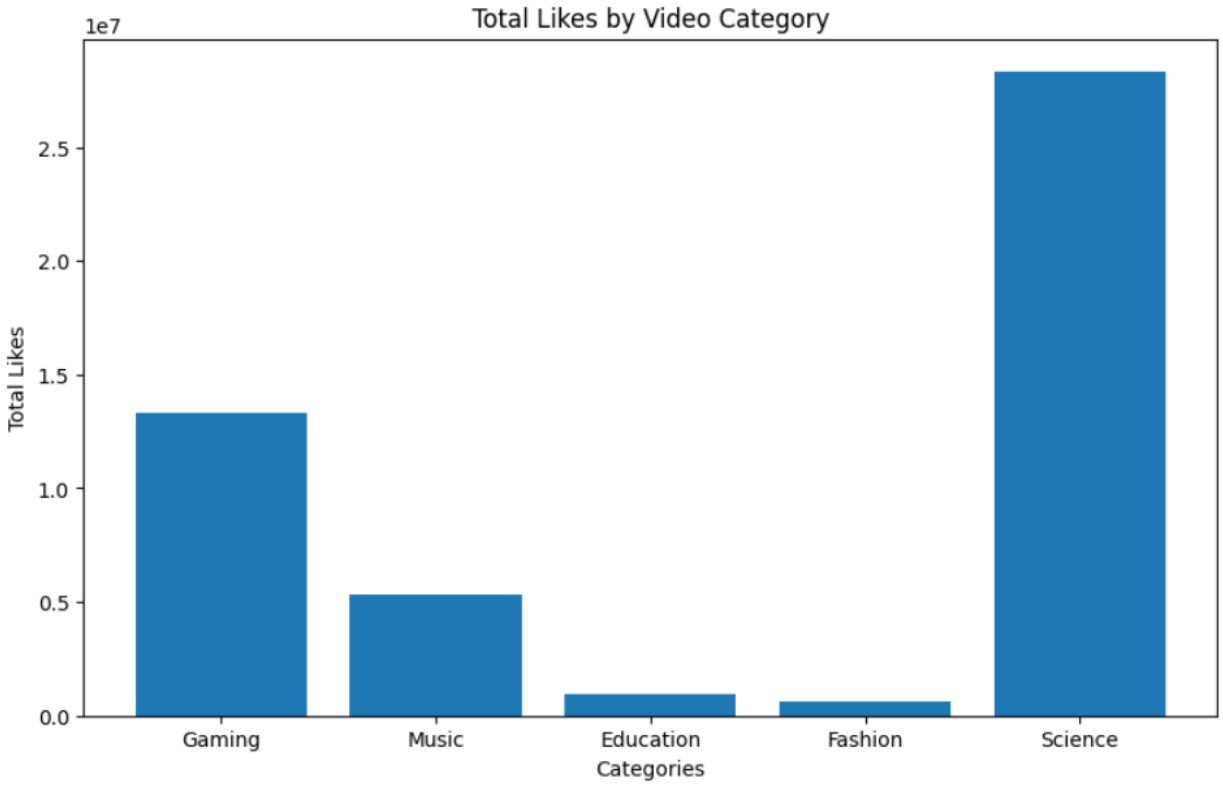
Во дијаграмот се видливи следните категории со соодветни бројки:

Science: Има најголем број, со околу 140 видеа.

Gaming: Има втор најголем број, со околу 110 видеа.

Music: Има значително помал број, со околу 60 видеа.

Fashion и Education: Имаат уште помал број, со околу 20 видеа.



Слика 2

На втората слика е прикажан дијаграм кој го визуелизира вкупниот број на лајкови за различните категории на видеа. Резултатите покажуваат дека категоријата „Science“ има најголем вкупен број на лајкови, што укажува на висока популарност и посветеност на публиката со оваа категорија.

Тука се видливи следните категории со соодветни вкупни лајкови:

Science: Има најголем вкупен број на лајкови, со околу 28 милиони.

Gaming: Има исто така поголем вкупен број на лајкови, со околу 15 милиони.

Music: Има околу 7 милиони вкупни лајкови.

Education: Има околу 2 милиони вкупни лајкови.

Fashion: Има најмал број на лајкови односно околу 1 милион вкупни лајкови.

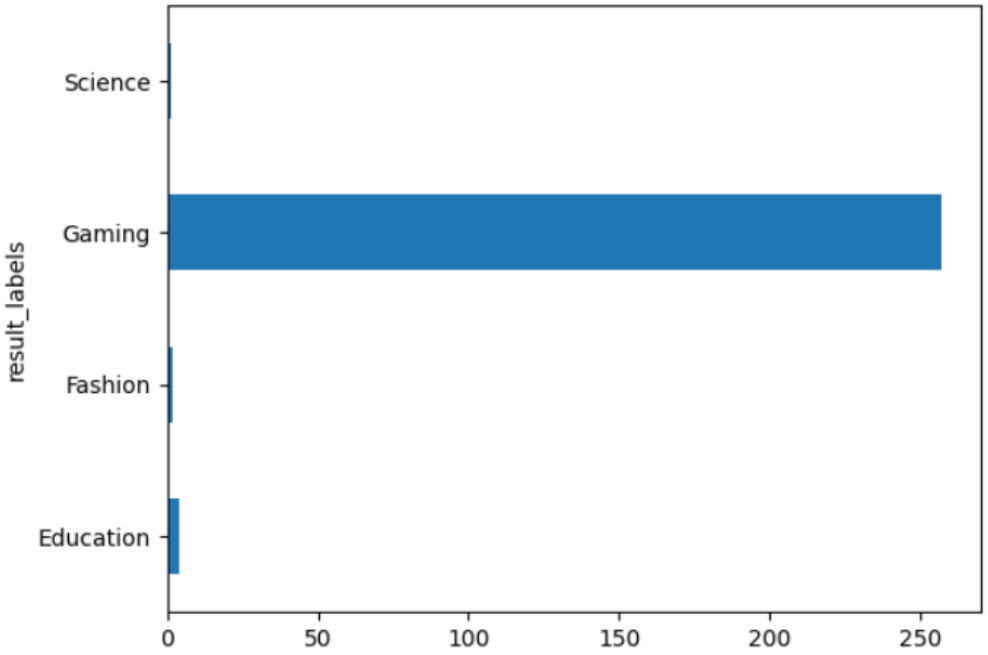
5.2 Заклучок

Од резултатите на слика 1 и слика 2 може да се заклучи дека категоријата „Science“ доминира не само по бројот на видеа, туку и по вкупниот број на лајкови. Ова покажува дека гледачите имаат најмногу интеракција со оваа категорија.

Категоријата „Gaming“ е помалку застапена, но сепак има значителен број на видеа што укажува на интерес на публиката. Категоријата „Education“ и “Fashion” имаат најмалку видеа а и најмал број на лајкови, што може да значи дека има помал интерес за создавање на образовни и модни видеа а воедно и содржината е помалку популарна на овој канал за гледачите.

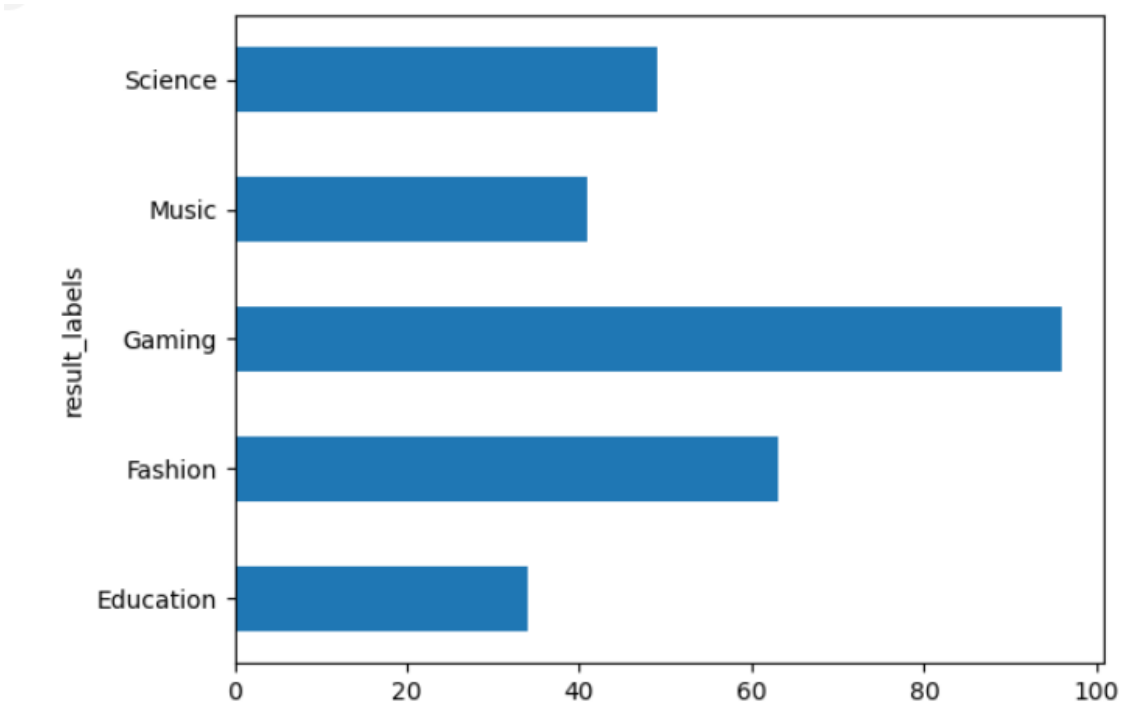
Со помош на трансформерот успешно се класифицирани видеата а и воедно се пресметани и вкупниот број на лајкови за истите. Овие резултати помагаат да се разбере фокусот на содржините и можат да се користат за идентификување на можностите за проширување и подобрување на каналот. Популарните категории како „Gaming“ и “Science” во овој случај може да продолжи да привлекува голема публика а со тоа и да расте каналот.

6. ДРУГИ АНАЛИЗИ



Слика 3

Покрај овој канал се анализираше и каналот disruptive builds кој е насочен само на играње игри. Со истата постапка како и за другиот канал како што може да се види на слика 3 се добиваат резултати дека скоро 250 од 264 видеа се насочени на играње игри додека многу помал дел се поделени во останатите класи, во класата музика пак нема ниту едно видео класифицирано. Од причина што каналот е насочен на играње игри, резултатите се очекувани.



Слика 4

За крај е земен канал именуван како emma chamberlain кој е всушност личност која објавува видеа од секојдневието и не е точно насочен кон одреден тип на содржина. Може да се забележи дека креаторот на каналот најмногу снима видеа кои се поврзани со игри а потоа следуваат fashion и music. Ако навистина се следат видеата на овој канал, ќе се забележи дека нејзината содржина е распределена и нема фокус кон само една содржина и еден тип на видеа кои ги објавува.

7. РЕФЕРЕНЦИ

The Slo Mo Guys, disruptive builds и emma chamberlain - youtube канали преку коиии се обработуваат и анализираат податоци - https://www.youtube.com/@theslowmoguys
https://www.youtube.com/@disruptivebuilds
https://www.youtube.com/@emmachamberlain

Hugging Face, платформа преку која е земен трениранот модел - https://huggingface.co/